Universitat de Lleida

TREBALL FINAL DE GRAU

Estudiant: POL GÒDIA SOLÉ

Titulació: Doble Grau en Enginyeria Informàtica i ADE

Títol de Treball Final de Grau: **FINANCIAL BANKRUPTCY PREDICTION WITH THE USE OF ARTIFICIAL INTELLIGENCE**

Director/a: **Xavier Sabi Marcano i Carlos Ansótegui Gil**

Presentació

Mes: Juny

Any: 2021

Acknowledgements

I would like to thank all the people that helped me during this project. From the University, I would like to thank my director, Xavier Sabi, for all the support and advice that he has given to me during these months, as well as Carlos Ansótegui, for his help in the Artificial Intelligence field.

I also want to thank all my relatives, who have always supported me, specially my parents, my sister, and my girlfriend, who have borne all my explanations about the project without exactly understanding what I was talking about.

Contents

1 Introduction **4**
- 1.1 Objectives . . . 4
- 1.2 Document structure . . . 5

2 State of the art **6**
- 2.1 Previous investigation in insolvency prediction . . . 6
- 2.2 Previous models . . . 7
  - 2.2.1 Non-centered Models . . . 7
  - 2.2.2 Centered Models . . . 9
  - 2.2.3 Previous models with the use of artificial intelligence . . . 10
- 2.3 Construction Sector . . . 11

3 Model **12**
- 3.1 Financial model . . . 13
- 3.2 Variables selection . . . 14
- 3.3 Classification of variables . . . 16

4 Data Treatment **18**
- 4.1 Data Selection . . . 18
- 4.2 Data Cleaning . . . 21
- 4.3 Sample . . . 22
  - 4.3.1 Final Attributes . . . 22
  - 4.3.2 Final Sample . . . 23

5 Experimental Evaluation **26**
- 5.1 Software used . . . 26
  - 5.1.1 Pandas . . . 26
  - 5.1.2 Numpy . . . 26
  - 5.1.3 Scikit-learn . . . 26
- 5.2 Preprocessing . . . 26
- 5.3 First approach on validating the model . . . 28
  - 5.3.1 Algorithms used . . . 28
- 5.4 Final Results . . . 31

6 Conclusions and future work **38**

List of Figures

2.1 Importance of construction sector to Spanish PIB 11

4.1 Evolution of construction subsectors 25

5.1 Cross-validation process 27

List of Tables

3.1 Financial/Non-financial variables classification 17
3.2 Classification of financial variables 17

4.1 Final variables of the model 23
4.2 Sample summary 24

5.1 Final results resume year before 37
5.2 Final results resume year before 37

Chapter 1

Introduction

The study of business failure has been a focal point of the financial literature during the last decades. This effort has led to a wide variety of prediction models, supported by many different methodologies. Unfortunately, the recent crises have increased the business failure during the last years.

Correct utilization of the financial information could have anticipated failures that occurred. For this reason, the main goal of this project is to create a business failure prediction model, to know if it is possible to predict financial failure 2 years in advance. To conclude if the model can be used for insolvency prediction, it is going to be validated through artificial intelligence techniques.

The project starts with an investigation of the state of the art, analysing past insolvency prediction models and their results. After this, the different sectors and databases where data could be gathered were analysed, in order to see the viability of the project.

The weight that the construction sector has in the Spanish economy justifies the attention in it, and that the model is centered in this sector.

Once we reached this point, the objective was to create an insolvency prediction model, collecting information from other models that had succeeded and adding personal considerations to it. Cleaning and treating all the data has been the step before evaluating the model with the use of different artificial intelligence techniques through the use of the software machine learning library for the Python programming language, Scikit-learn.

The final results obtained will tell us if using the model created and the data obtained, insolvency prediction could be predicted two years before happening. If these results are positive, the model could have other applications in the field.

1.1 Objectives

The main goal of this project is to confirm if it is possible to predict the bankruptcy failure of a company with its financial and non-financial information 2 years in advance. To achieve this, we will need to accomplish the following objectives:

- Previous investigation about the state of the art.
- Select the different variables that will be included in the model.
- Gather the correct information in order to be related with the model previously selected.

- Treat the information collected, transform it into valid data, and validate it after with the appropriate software.
- Use Machine Learning (ML) algorithms to prove the efficiency of the model.

1.2 Document structure

This document is structured in 6 chapters:

Chapter 1: Introduction

Corresponds to the introduction of this project.

Chapter 2: State of the art

Presents the state of the art of financial insolvency, previous investigation and results, as well as the sectors that have been more affected.

Chapter 3: Model

Describes the different financial and non-financial variables that have been selected and its reason.

Chapter 4: Data Treatment

Explains all the process that starts from the data selection in "Sistema de Análisi de Sistemas Ibéricos (SABI)" database, and its following process of data cleaning.

Chapter 5: Experimental evaluation

This chapter describes the preprocessing of the classes and the different algorithms used for estimating the model.

Chapter 6: Conclusions and future work

Presents the conclusions of this project and the future research work.

Chapter 2

State of the art

In this chapter, we are going to resume all the previous investigations done, centering on the financial insolvency prediction state of the art, its previous studies, and results.

Companies' bankruptcy has been a huge problem since the firsts economic crisis appeared. The study of the reasons why companies failed has led to a wide variety of analyses, intending to have the capacity to predict these bankruptcies.

2.1 Previous investigation in insolvency prediction

There have been a lot of investigation studies oriented to determine the factors that have caused the business failure, with an especial concern on the incidence of predicting it before it happens.

The analysis of financial insolvency has its beginnings in the United States during the sixties, with Beaver (1966) [Beaver, 1966] and Altman (1968) [Altman, 1968]. It was not till the eighties when it started the research in Spain, mainly due to the crisis of banks and insurance companies.

The first studies were done with individual analysis of the variables. Altman (1968) was the first to use multi-discriminate analysis (MDA), building the famous Z-score model, which nowadays continues to be used in many studies on the field.

The start of the application of artificial intelligence techniques was in the nineties, where neuronal networks began to be introduced in insolvency prediction investigations, through Tam (1991) and Tam and Kiang (1992) [Tam and Kiang, 1992].

Another artificial intelligence method used was support vector machines (SVM) with Shin et Al.(2005) [Shin et al., 2005] and Min and Lee (2005) [Min and Lee, 2005], concluding that its efficiency was better than MDA, Logit, and neuronal networks.

Focusing on the model, we could also differentiate between non-centered (global) and centered models. Centered models study a determined sector or feature. In our case, the decision has been to focus the model on the construction sector, as it is one of the more affected sectors in Spain during the last years. An analysis of the previous literature proves that the number of non-centered models in comparison with centered ones is elevated. The main problem was that in the past, the databases did not have enough information to center all the analysis in a particular sector.

On the one hand, statistical and computational methods have conditioned the development of the literature about insolvency prediction. However, it exists other relevant criteria to understand the evolution of the investigation done, like the number of ratios and variables used and the evolution of the results obtained.

On the other hand, the number of ratios used in the different models has a wide range, having from only 1 ratio to a total of 57 in the same model. From this ratios, the majority of them are financial, and a lot of previous studies do not take into account non-financial variables in their models. Here, we find the first decision to take in our model. After researching information from previous studies, the average number of variables used is between 8 and 10, and the selection of them is essential to evaluate the model [Jodi L. Bellovary and Akers, 2007].

According to Bellovary et Al. (2007), during the sixties, the precision range was between 79% and 92%. During that decades, the predominant models were based on Univariate Discriminant Analysis. The majority of the studies done have used information of a year before the failure, obtaining an average level of prediction of 81%. From the results obtained, it is noticed that the capacity of the models to predict, diminishes considerably when information from more than a year before the failure is used.

In the area of our study, we will consider the concept of financial insolvency when an arrangement or pre-arrangement with creditors it is produced, which it means that the company cannot deal regularly with the obligations it has. A company is it considered to be insolvent when it has the legal status of bankruptcy, according to the considerations done by "Ley Concursal 22/2003 de 9 de julio"[1], as well as the modifications made by "Real Decreto Ley 3/2009 de 27 de marzo"[2] for urgent measures due to the evolution of the economic situation. [González and y A. Y. Sánchez, 2003]

According to the legislation mentioned, it is defined the "Concurso de acreedores" as the legal procedure that starts when a natural or legal person becomes in an insolvency situation where it is not able to face the totality of the payments it owes. In the 22/2003 Law, it mentions the situations of suspension of payments and bankruptcy, considering the suspension of payments as transient insolvency and the bankruptcy as definitive insolvency, both referred to legal people.

2.2 Previous models

The failure prediction has been studied in many different ways, but there are some studies that precede a big amount of them.

Generally, models that had been developed traditionally for insolvency prediction, were formed by samples of medium and big companies, which belonged to the industrial and commercial sectors in a wide sense.

As we have mentioned before, models can be non-centered and done for every type of company, or centered by a sector or type of company. There is not a definitive conclusion about the superiority of centered or non-centered models [Jodi L. Bellovary and Akers, 2007]. It is possible that the lack of a conclusion is since it has not been possible to compare the models in a homogeneity way, because of the disparity of methodologies, approaches, databases, temporal periods, and countries, among other things. [Laguillo et al., 2017].

2.2.1 Non-centered Models

As explained before, non-centered models are those that have been elaborated with heterogeneous samples, meaning the companies are not from the same economic sector.

[1] https://www.boe.es/buscar/doc.php?id=BOE-A-2009-5311
[2] https://www.boe.es/eli/es/rdl/2009/03/27/3

In the non-centered model's literature, we can highlight the studies of Beaver (1966), Altman (1968),Deakin (1972), Wilcox (1973), Ohlson (1980), Gentry et al. (1985), Coats y Fant (1992), Fletcher y Goss (1993), Altman (1994), Chang-Lee et al. (1996), Yi-Chung et al. (2005), Pindado et al. (2008), Chen et al. (2011), Kwak y Gang Kou (2012), Sangjae y Wu (2013), summarized in an article with studies from 1930 to present. [Manuel Rodríguez López and de Llano Monelos, 2006].

After carrying out the previous research, the three models that have been used more by observation are the ones that follow:

Beaver 1966

Beaver (1996), in his ground-breaking work, used Moody's Industrial Database, which contained financial information about listed companies in the stock market.

He did a uni-variant study of 30 ratios, on which concluded that the more precise ones were:

- Net Profit / Total Debt
- Net Profit / Sales
- Net Profit / Net Worth
- Cash Flow / Total Debt
- Cash Flow / Total Assets

Altman 1968

Altman (1968), considered the classic study par excellence, used the Discriminant Analysis and applied it to a group of industrial companies.

Altman z-Score = 1,2 T1 + 1,4 T2 + 3,3 T3 + 0,6 T4 + 1T5

- T1 = Working Capital / Total Assets
- T2 = Retained Profits / Total Assets
- T3 = EBITDA / Total Assets
- T4 = Market Capitalization / Total Debt
- T5 = Sales / Total Assets

This model obtained a 95% prediction the year before the failure, but plunged to a 72% two years before the failure.

Bellovary 2007

Bellovary et al. (2007) reviewed more than 150 bankruptcy studies published from 1965 to 2004. The number of financial variables used in models varied from 1 to 57 with an average equal to 10. In total 752 variables were used in the reviewed studies. In the case of models with the highest classification accuracy, the number of variables ranged from 2 to 21. Bellovary et al. (2007) concluded that a higher number of variables included in the model is not related to higher prediction accuracy due to multicollinearity [Altman, 2015].

According to Bellovary et Al. (2007), 752 different variables have been used in studies related to insolvency prediction. The most used variables have been:

- 1. Net Profit / Total Assets, which is included in 54 studies
- 2. Current Assets / Current Liabilities
- 3. Working Capital / Total Assets
- 4. Retained Profits / Total Assets
- 5. Pre-tax profit / Total Assets
- 6. Sales / Total Assets
- 7. Total Debt / Total Assets
- 8. Current Assets / Total Assets
- 9. Net Profit / Net Worth

The availability of information and valid registers was a limiting factor for the development of more centered or specific models. For this reason, till more databases with complete information were deployed, centered models represented a small percentage.

2.2.2 Centered Models

In this section, we are going to examine those studies about insolvency prediction that have been based on a specific economic sector.

The most popular of the centered models are the ones used for credit entities (Santomero and Vinso, 1977 [Santomero and Vinso, 1977]; Martin-del-Brio and Serrano-Cinca, 1995 [Serrano-Cinca and Martín-del Brío, 1993]). Others that have also been popular are centered on industrial companies (Altman, 1968 [Altman, 1968]; Appetiti, 1984 [Peres and Antão, 2017]; Zavgren, 1985 [Zavgren, 2006]).

Recently, there have also appeared studies in companies of different sectors, like Internet (Wang,2004 [Wang et al., 2004]), hostelry (Park y Hancer, 2012 [Park and Hancer, 2012]; Fernández, Cisneros y Callejón, 2016 [Fernández-Gámez et al., 2016]), agriculture (Mateos-Ronco et al., 2011 [Mateos-Ronco and Mas, 2011]), construction (Gill de Albornoz y Giner, 2013 [Gil de Albornoz and B.Giner, 2013]), and services (Keener, 2013 [Pang, 2013]).

Langford et al. (1993) [Shi and Li, 2019] examined two models to value the financial viability of the construction companies. They concluded that the techniques used were valuable but they needed to be used with financial data from other companies of the construction sector. The article recommended a specific Z model for construction companies.

Abidali y Harris (1995) [Abidali and F.Harris, 1995] worked on the development of a system to identify construction companies near to bankruptcy. The system was composed of a predictive discriminant that used financial ratios, giving it a value Z. The model was estimated with 7 variables (4 financial ratios and 3 tendency variables). The Z model obtained a 90% accuracy in the non-bankrupted companies and a 100% from the companies that failed.

Minguez (2006) [Mínguez-Conde, 2006] gave empiric evidence about the insolvency prediction in construction, as there were not previous specific models for this sector. He estimated through Logit and Cox models. Using SABI Database, selected 126 companies, where 63 failed and the other 63 no. The results from Logit model were 3,85% error type I and 19,23% error type II, with data from the year before.

Stroe and Barbuta-Misu (2010) [Stroe and Bărbuţă-Mişu, 2010] predicted the financial result of a company in Rumania., getting the information of 11 construction companies. Using the Z model they were able to predict bankruptcy with a rate of 81,82%.

Gil de Albornoz y Giner (2013) [Gil de Albornoz and B.Giner, 2013] investigated if the estimation of specific models could predict better business failure than a generic one. They used 4.600 Spanish companies, where almost a half were from the construction sector. After applying different ratios, they conclude that the sectoral estimation allows a better classification of the failed companies than the general estimation. Otherwise, it classified worse the non-failed companies. To sum up, the results confirmed that profitability, stock rotation, indebtedness, liquidity, age, and general economic conditions affected the probability of insolvency, but existing sectoral differences.

2.2.3 Previous models with the use of artificial intelligence

The appearance of insolvency prediction models with the use of artificial intelligence started in the nineties, with the main use of Neuronal Networks (NN).

The main previous studies in this field are Bell et al., 1990; Tam y Kiang, 1992; Serrano y Martín, 1993; Wilson y Sharda, 1994; Altman et al.,1994; Koh y Tan, 1999; Yang et al., 1999; Brockett et al., 2006; Tsai, 2008; Boyacioglu et al., 2008; Kim, 2011; Xiaosi et al., 2011, according to a revision of previous studies [Manuel Rodríguez López and de Llano Monelos, 2006].

Another of the pioneer works in the introduction of NN was Coats and Fant (1992) [Abid and Zouari, 2000]. The objective was to describe what NN were and their application to the insolvency prediction field.

To appreciate the classification power of NN, took a sample of financially stressed companies, with the financial data of the 3 previous years before reaching that situation. After that selection, he divided the samples into training and testing.

NN were able to predict correctly with an accuracy of 91% and 96% the companies that had financial problems and the ones that not, respectively, whereas MDA only had an accuracy of 72% from the failed ones and 89% the non-failed.

Fletcher and Goss (1993) [Fletcher and Goss, 1993] illustrated the development of a business insolvency prediction model using a particular class of NN, named Retropropagation. The NN model had a better level of prediction, with an 82% rate against a 71% rate obtained through Logit.

Wilson and Sharda (1994) [Wilson and Sharda, 1994] realized a comparative analysis of the predictive capacity of NN against MDA. The obtained results showed that the rate of NN was slightly superior to MDA.

What is criticized in the performance of NN is that it works as a black box, which is not the most appropriate to describe the financial dysfunctions neither providing guides to avoid failure.

Chan et Al. (2011) [Chan et al., 2005] proposed the Support Vector Machine (SVM) methodology to predict insolvency in German companies. The main goal was to demonstrate that a well-specified SVM model could obtain better results than MDA and Logit. The final results confirmed that SVM was able to predict insolvency 10% better than the other methodologies mentioned.

The named artificial intelligence classifiers like Neuronal Networks (RNA), Support Vector Machine (SVM), Evolution Algorithms (EA), Rough Set (RS), and Decision Trees (DT) have been used in the application of insolvency prediction and are supported by remarkable results.

They offer the advantage of not being subject to the assumptions demanded by statistical techniques.

2.3 Construction Sector

The weight of the construction sector has a huge socioeconomic significance, for its contribution to the "Producte Interior Brut (PIB)" and the workplaces that it has created directly or indirectly.

On the 1st of January 2019, the number of active companies in Spain was 3,36M, where the construction sector represented 12,6% of the total in 2019 [INE, 2019].

For these reasons and as the construction sector had more information about companies that had failed, we decided to create a construction-centered model.

In Figure 2.1, we can see the importance that the construction sector has to the Spanish PIB, noticing a big reduction after the crisis that started in 2007.

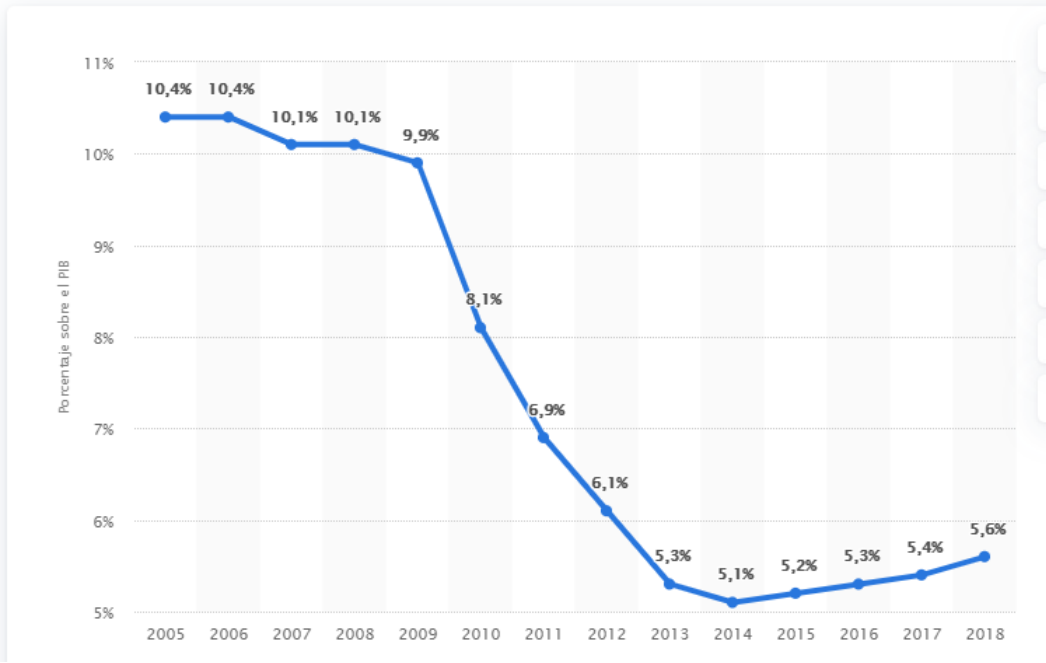

Figure 2.1: Importance of construction sector to Spanish PIB
Source: Statista (2019). Peso del sector de la construcción en el pib de españa 2005-2019

Chapter 3

Model

After analysing previous models we decided to center our study on the Spanish construction sector, to be more accurate with the results and the data collected.

The general improvement of the model accuracy should be linked to the selection of ratios. The question arises of how to select ratios properly and how many ratios should be used in a model.

Balcaen and Ooghe (2006) [Balcaen and Ooghe, 2006] reviewed business failure studies over 35 last years and concluded that there is little consensus on which variables are the best in discriminating between failed and non-failed firms.

Balcaen and Ooghe underlined the lack of theoretical framework for variable selection although some researchers have tried to base selection on, for example, the cash-flow theory, the gambler's ruin model, option pricing, or the integrated ratio model. In reality, variables selected for a given failure prediction model are often sample and environment-specific and the results are difficult to generalize. Karels and Prakash (1987) [Karels and Prakash, 2006] emphasized that careful selection of variables is necessary to improve the performance of the models. In the same way, Zavgren and Friedman (1988) [Zavgren and Friedman, 1987] indicated three drawbacks of previous studies, one of those being an arbitrary selection of variables. Balcaen and Ooghe (2006) concluded their review arguing that simple models with a small number of variables may gain significantly in classification accuracy in comparison to complicated models, due to the 80/20 Pareto rule, and the law of diminishing returns.

The causes of bankruptcy are numerous, as many authors have pointed out (Argenti, 1976; Lussier, 1995; Blazy and Combier, 1997; Sullivan et al., 1998; Bradley, 2004). The typology suggested by Blazy and Combier (1997) [Davydenko and Franks, 2008] offers a relevant synthesis of these major causes:

- accidental causes: malfeasance, death of the leader, fraud, disasters, litigation;
- market problems: loss of market share, failure of customers, inadequate products;
- financial threats: under-capitalization, cost of capital, default on payment, loan refusal;
- information and managerial problems: incompetency, prices and stocks, inadequate organization;
- macroeconomic factors of fragility: declining demand, increased competition, credit rationing, high-interest rates;

- costs and production structure: excessive labour costs, over-or under-investment, sudden loss of a supplier, inadequate production process;
- strategy: failures of major projects, acceptance of unprofitable markets.

Financial problems, then, have many causes, and these causes are easily determined. But finding the variables that may reflect these factors is altogether different [du Jardin, 2009].

3.1 Financial model

Financial modeling is the design and planning of a financial model, that is used when preparing a business plan or shaping the financial structure of a company. The purpose is to capture the reality of the company that provides a transparent image of the global situation of the company. A good financial model is detailed enough not to leave out any key factors but no so complicated that it cannot be used in any situation. The basis of the financial model reflects the Balance Sheet, the Profit and Loss Account, and the Cash Flow control.

Financial models can become very complex and therefore the data and its visualization must be carefully studied. A model that is useless due to its complexity leads to errors as an incomplete or unrealistic model.

After carrying out all the previous research, we observed that the number of companies used in the samples was small and that if we could obtain a higher number of them, the validation of the model would be more effective.

The chosen ratios are based on the previous literature that we had read, selecting those that have obtained better results in the past and selecting others that we have considered that may have an important impact on the results. The ratios are based mostly on margin, rotation, solvency, debt and structure.

The most frequent sources of insolvency concerning the firm's financial decision-making are the debt-equity ratio, lack of own financial reserves, problems with the enforceability of claims and financial inflexibility in response to the decline in sales [Altman, 2015].

For choosing the attributes of the model, we have taken into account the following points, which have been considered important when evaluating the financial situation of a company.

- **Profitability**: This dimension associates the return of a determined investment. As higher the profitability it is, the better perspective of the financial situation.
- **Liquidity**: This dimension is based on the capacity that a company has to face the short-term obligations. Is the facility in which an asset can be converted into cash.
- **Solvency**: Liquidity on a long-term. It measures the capacity that the company has to face short and long-term obligations.
- **Indebtedness**: is the dimension that relates the total amount of debt that the company has with the resources it has. The level of debt is higher on companies that require external sources of financing.
- **Structure**: This is the way that the assets and liabilities are divided, according to if they are short-term or long-term.
- **Margin**: This is the difference between the profitability of a financial product and the cost that this product had.

- **Rotation**: It measures the recuperation rate of the operational investments through the sales revenues. It is the number of times that during the financial year, the company covers the value of its investment through the revenue.

3.2 Variables selection

After conduction the previous investigation and considerations, the selection of the ratios was going to play a crucial role in the project. For this reason, we based the selection on popular ratios in previous studies and our ideas, to create a balanced model.

In terms of variable selection, the majority of studies only included financial variables, but we thought that non-financial variables could also play an important role.

One of the first studies that included non-financial variables for bankruptcy prediction of UK small firms was delivered by Keasey and Watson (1987) [Keasey and Watson, 2006], who tested Argenti's hypotheses (1976). Non-financial variables included 1 company age variable, 4 managerial structure variables (changes of members, potential autocratic regime), 8 variables related to the potential to "cook the books", such as delay in submission (4), qualified audits (3) ,and change of auditor (1), 3 variables showing leverage, 2 variables on management accounting system and 1 age variable. Non-financial variables were supplemented by 28 financial ratios. Keasey and Watson (1987) presented 3 models based on financial variables only, non-financial variables only, and both (mixed or combined model) with conclusions that marginally better predictions may be observed with the use of non-financial variables together with financial ones.

Laitinen (1999) [Laitinen and Kankaanpaa, 1999] used, for credit risk estimation of Finnish companies, 35 variables, within 16 non-financial variables related to the age of the firm (1), payment behavior of the company (3), management structure and their financial situation and payment behavior (8), industry (2), number of enquiries about the firm in credit information bureau (1) and legal form (1).

For German data, Grunert et al. (2005) [Grunert et al., 2005] showed the importance of non-financial variables. Besides six financial variables, they used management quality and market position as non-financial variables. In conclusion, they pointed out that the combined use of both types of variables provided better default prediction. Altman and Sabato (2007) [Altman and Sabato, 2007] pointed out that model prediction accuracy may be improved by the use of qualitative variables. They divided non-financial variables into four categories: type and sector, size and age, reporting and compliance, and operational risk (including auditor's opinion and country court judgment).

Wilson et al. (2013) [Wilson, 2013] analyzed family vs. non-family business survival and the role of boards of directors. Those two factors should be treated as non-financial variables. Family business was defined by the ownership structure and also with the percent of shares owned by directors. Board composition was analyzed from the perspective of instability (resignations 2 years before bankruptcy), gender structure (presence of female dummy), age (average age and age variation), experience (average day experience in the firm and in the sector), living in the same county, ratio of directors that failed in the past, number of multiple directorships and a ratio of independent directors. They combined board-related variables with financial and non-financial variables used by Altman et al. (2010).

The fact that we selected the construction sector did not play a key role in the selection process, but in the analysis of the final data obtained.

Finally, the financial and non-financial ratios selected are the following.

- **365/Stock Rotation**: This ratio measures the number of times that the stocks are converted into cash, depending on the type of business.

- **Collection Period**: This is the time that passes between we sell a product and the client pays us. As lower is this value, it will mean that we are receiving the money sooner.

- **Suppliers Payment Period**: This is the time that passes between we buy a product and we pay it to the supplier. As higher is this value, it will mean that we have more time to pay, so more cash we will have during that period.

- **Audit**: An audit is a verification to know if the audited company is working with the internal and external regulations. If the results of the audit are favourable or not will determine an indicator about the financial accounts of the company.

- **Number of Companies in Corporate Group**: A company that is supported by a corporate group has more possibilities to be healthy. Or oppositely, if a company of the group fails it could fail, too.

- **Legal Form**: This is the legal identity that a company has, depending on its individual, corporate or cooperative character.

- **Share Capital**: This is the money that shareholders invest to start or expand the business. As higher this value it is, the more resources the company will have.

- **CNAE** It is "Clasificación Nacional de Actividades Económicas" and assigns a code to each economic activity. It is useful to know the exact activity that a company does.

- **Operating Result/Operating Income**: It measures the margin of the company. As higher is this ratio, better information about the company.

- **Sales / Total Assets** : It is the number of times that the sales revenue covers the assets of the company. It is a measure of the efficiency that the company obtains from their assets during a determined period of time. In the construction sector, the ratio is between 0,5-0,62.

- **Current Assets / Current Liabilities**: Second most used ratio according to Bellovary et Al, in a total of 51 studies. It talks about the short-term liquidity of the company, comparing the current assets with the current liabilities. It shows the capacity that the company has to face the obligations that have a short-term expiration. If the result is lower than 1, it will mean that we do not have sufficient resources to face the obligations. According to the obtained information, the ratio in this sector is about 1,35-2 in an equilibrium situation. It also allows us to know whether the rolling fund is big or not.

- **Current Assets / Total Assets**: Ratio that allows us to know the assets structure of the company, knowing the % of current assets. In the construction sector, we are going to find more short-term assets than long-term, as they usually rent their equipment instead of buying it. According to ratios realised in companies of the sector, the assets are divided 30-70. Is one of the most used top 10 ratios according to Bellovary et Al (2007) [Jodi L. Bellovary and Akers, 2007].

- **Current Liabilities / Total liabilities**: It shows which part of the debt is short-term and which is long-term. We say that the indebtedness has quality when the major part of the debt is long-term so that the ratio should be as closer to 0 as possible.
- **Financial Debt / Cash Flow**: This ratio shows the capacity that the company has to return the financial debt that it has according to the cash flow it generates. The goal is to have the ratio as smaller as possible.
- **Total Assets / Total Debt** This ratio is going to inform about the indebtedness of the company, and the capacity we have to cover it with the assets. If the value is lower than 1, it will mean that we have more obligations than resources, which means a bankruptcy situations.
- **Sales evolution** This measure is going to help us to see the evolution of the sales of a year comparing to the year before. A plunge of the sales could be a synonym of future problems.
- **Employees evolution** It is a percentage to know the variation that the number of employees has suffered. If this variation is positive, it will mean that the number of employees has increased, and if it has a negative variation, the number of employees would have decreased, what could be a negative fact for the company.
- **Size** : It is measured through the logarithm of the total assets. It will help us as a non-financial ratio. Bigger companies tend to have fewer financial problems than smaller ones.
- **Age**: From the constitution date till the day of gathering the data. Companies that have been created more recently, could have more financing problems.

3.3 Classification of variables

The different variables selected for the model can be divided in different ways. Firstly, we have distinguished between financial and non-financial variables, as we can see in Table 3.1 . Secondly, we have classified the financial ones according to what they represent, as we can see in Table 3.2.

	Financial	Non-financial
365/Stock Rotation	X	
Collection Period	X	
Suppliers Payment Period	X	
Audit		X
Number of Companies in Corporate Group		X
CNAE		X
Share Capital	X	
Operating Result/Operating Income	X	
Sales/Total Assets	X	
Current Assets/Current Liabilities	X	
Current Assets/Total Liabilities	X	
Current Liabilities/Total Liabilities	X	
Financial Debt/Cash Flow	X	
Total Assets/Total Debt	X	
Legal Form		X
Sales evolution		X
Employees evolution		X
Size		X
Age		X

Table 3.1: Financial/Non-financial variables classification
Source: Own elaboration

	Profitability	Solvency	Indebtedness	Structure	Margin	Rotation
365/Stock Rotation						X
Collection Period		X				X
Suppliers Payment Period		X				X
Share Capital				X		
Operating Result/Operating Income	X				X	
Sales/Total Assets						X
Current Assets/Current Liabilities		X				
Current Assets/Total Liabilities				X		
Current Liabilities/Total Liabilities				X		
Financial Debt/Cash Flow		X	X			
Total Assets/Total Debt		X	X			

Table 3.2: Classification of financial variables
Source: Own elaboration

Chapter 4

Data Treatment

In this chapter we are going to describe the transformation of the information collected from the database into data, giving it a concrete format. We will also describe the variables selected and the final sample.

4.1 Data Selection

The main goal of this project is to validate an insolvency prediction model and to achieve it, it is necessary to gather data related to it. We needed information about healthy firms and bankrupted ones, to analyse and compare them. In our case, we have selected two different samples.

All the information has been collected from the database "Sistema de Análisis de Balances Ibéricos (SABI)"[1], which provides information from many companies in Spain and Portugal. We have chosen this database because is the one that has more information about failed companies, which is essential for the research. It is also the database that the University has allowed access to it.

Once we selected the database, the filters that we used were the following: For the failed companies:

- -"CNAE 2009(Sólo códigos primarios): 41 - Construcción de edificios, 42 - Ingeniería civil, 43 - Actividades de construcción especializada "
- -"Incidencias: Reclamaciones administrativas, Incidencias judiciales"
- -"Estados España: Concurso, Suspensión de pagos, Quiebra"
- -"Fecha de constitución: hasta 31/12/2017"

For the healthy companies:

- -"CNAE 2009(Sólo códigos primarios): 41 - Construcción de edificios, 42 - Ingeniería civil, 43 - Actividades de construcción especializada"
- -"Estados España: Activa"
- -"Fecha de constitución: hasta 31/12/2017"

According to CNAE Codes:

[1] https://sabi.bvdinfo.com/version-202115/home.serv?product=SabiNeo

- 41: "Construcción"
  - 412: "Construcción de edificios"
    - 4121: "Construcción de edificios residenciales"
    - 4122: "Construcción de edificios no residenciales"
- 42: "Ingeniería Civil"
  - 421: "Construcción de carreteras y vías férreas, puentes y túneles"
    - 4211: "Construcción de carreteras y autopistas"
    - 4212: "Construcción de vías férreas de superfície y subterráneas"
    - 4213: "Construcción de puentes y túneles"
- 43: "Actividades de Construcción Especializada"
  - 431: "Demolición y preparación de terrenos"
    - 4311: "Demolición"
    - 4312: "Preparación de terrenos"
    - 4313: "Perforaciones y sondeos"
  - 432:"Instalaciones eléctricas, de fontanería y otras instalaciones en obras de construcción"
    - 4321: "Instalaciones eléctricas"
    - 4322: "Fontanería, instalaciones de sistemas de calefacción y aire acondicionado"
    - 4329: "Otras instalaciones en obras de construcción"
  - 433: "Acabado de edificios"
    - 4331.- "Revocamiento"
    - 4332.- "Instalación de carpintería"
    - 4333.- "Revestimiento de suelos y paredes"
    - 4334.- "Pintura y acristalamiento"
    - 4339.- "Otro acabado de edificios"
  - 439.- "Otras actividades de construcción especializada"
    - 4391.- "Construcción de cubiertas"
    - 4399.- "Otras actividades de construcción especializada n.c.o.p."

After applying the filters above, we obtained information from 8.900 companies that have failed or with difficulties and 15.000 entries from healthy companies.

The different information that we need of the companies to be able to calculate the different ratios we selected previously, was the following:

- - "Forma jurídica"
- - "Capital Social"
- - "Fecha constitución"

- - "Nº of copmanies in corporate group"
- - "CNAE"
- - "Número empleados Últ. año disp."
- -"Número empleados Año - 1"
- -"Número empleados Año - 2"
- -"Número empleados Año - 3"
- -"Inmovilizado mil EUR Últ. año disp."
- -"Inmovilizado mil EUR Año - 1"
- - "Activo circulante mil EUR Últ. año disp."
- - "Activo circulante mil EUR Año - 1"
- -"Total activo mil EUR Últ. año disp."
- -"Total activo mil EUR Año - 1"
- -"Pasivo fijo mil EUR Últ. año disp."
- -"Pasivo fijo mil EUR Año - 1"
- -"Pasivo líquido mil EUR Últ. año disp."
- -"Pasivo líquido mil EUR Año - 1"
- -"Deute financer mil EUR Últ. año disp."
- -"Deute financer mil EUR Año - 1"
- -"Ingresos de explotación mil EUR Últ. año disp."
- -"Ingresos de explotación mil EUR Año - 1"
- -"Importe neto Cifra de Ventas mil EUR Últ. año disp."
- -"Importe neto Cifra de Ventas mil EUR Año - 1"
- -"Importe neto Cifra de Ventas mil EUR Año - 2"
- -"Importe neto Cifra de Ventas mil EUR Año - 3"
- -"Resultado Explotación mil EUR Últ. año disp."
- -"Resultado Explotación mil EUR Año - 1"
- -"Cash flow mil EUR Últ. año disp."
- -"Cash flow mil EUR Año - 1"
- -"Período de cobro (días) Últ. año disp."
- -"Período de cobro (días) Año - 1"

- -"Período de crédito (días) Últ. año disp."
- -"Período de crédito (días) Año - 1"
- -"Calificación auditor Últ. año disp."
- -"Calificación auditor Año - 1"
- -"Rotación de las existencias % Últ. año disp."
- -"Rotación de las existencias % Año - 1"

4.2 Data Cleaning

When we firstly collected all the necessary information to create the variables of the model, which we mentioned in Chapter 3, we exported it to Excel to analyse and clean it by applying different filters. The total amount of failed companies that we had information about was 8900 and 15000 for healthy companies. This unbalance is because exist more companies with good health than the ones that failed.

After analysing the information, we realised that the majority of the companies did not have all the variables needed, so we applied filters in order to transform that information into valid data.

The process of data cleaning done has been the following:

- -"Nº of companies in corporate group": Eliminated rows with value "Vacías"
- - "Número empleados Últ. año disp. y "Número empleados Año - 1": We have eliminated all the rows with value "n.d".
- -"Inmovilizado mil EUR Últ. año disp." and "Inmovilizado mil EUR Año - 1": We have eliminated rows with value "n.d".
- -"Activo circulante mil EUR Últ. año disp." and "Activo circulante mil EUR Año - 1": We have eliminated rows with value "n.d"
- -"Total activo mil EUR Últ. año disp." and"Total activo mil EUR Año - 1": We have eliminated rows with value "0" and "n.d",because a company that does not have assets represents that does not have an activity.
- -"Pasivo fijo mil EUR Últ. año disp.", "Pasivo fijo mil EUR Año - 1","Pasivo líquido mil EUR Últ. año disp." and "Pasivo líquido mil EUR Año - 1": We have eliminated rows with value "n.d"
- -"Ingresos de explotación mil EUR Últ. año disp." and "Ingresos de explotación mil EUR Año - 1" : We have eliminated rows with value "n.d"
- -"Importe neto Cifra de Ventas mil EUR Últ. año disp." and "Importe neto Cifra de Ventas mil EUR Año - 1": We have eliminated rows with values "0" and "n.d", because the ratios do not make sense with value 0.
- -"Resultado Explotación mil EUR Últ. año disp." and "Resultado Explotación mil EUR Año - 1": We have eliminated rows with value "n.d"

- -"Cash flow mil EUR Últ. año disp." and "Cash flow mil EUR Año - 1" : We have eliminated rows with value "n.d"
- -Período de cobro (días) Últ. año disp., Período de cobro (días) Año - 1, Período de crédito (días) Últ. año disp. and Período de crédito (días) Año - 1 : We have eliminated rows with value "n.d"
- -"Calificación auditor Últ. año disp." and "Calificación auditor Año - 1": We have eliminated rows with values "Vacías"
- -"Rotación de las existencias % Últ. año disp." and "Rotación de las existencias % Año - 1": We have eliminated rows with values "n.s" and "n.d".

Another modification of the data has been done in the variable Financial Debt/Cash Flow. We obviate this ratio when it turns negative because of a negative cash flow. For this reason, we changed all the negative values to 1.000, in order to differentiate clearly from the other values. We assigned the value 1.000 because the highest value for the rest of the companies was 487.

After cleaning the information, we obtained complete data about 1208 companies out of the initial 8900 from failed companies and 4410 out of 15000 from the non-failed companies.

In order to obtain more information about failed companies and to balance the two classes, we started a process to analyse the missing data and see if it can be replaced by the mean of the values from the other companies.

After analysing again the data and intending to balance the classes, we reached the following conclusion:

Different attributes have a high percentage of missing data and filling in the missing data with the mean of the other values could bring down the quality of the information collected.

The variables ""Número empleados Últ. año disp." and "Número empleados Año - 1" have approximately 40% of the information missing, as well as the variables "Deute financer mil EUR Últ. año disp" and "Deute financer mil EUR Año - 1". Once we have filtered it, the variables "Pasivo fijo mil EUR Últ. año disp." and "Pasivo fijo mil EUR Año - 1" have about 20% of missing data. Almost the same amount of missing data have the variables "Calificación auditor Últ. año disp." and "Calificación auditor Últ. año disp.". Another variable that has a big amount of missing data is "Rotación de las existencias % Últ.año disp." and "Rotación de las existencias % Año -1"

After eliminating this missing data, the maximum number of rows that we can obtain is almost the same that we had in the first approach to the data collected. For this reason, we decided to continue with the samples obtained, having to balance them.

4.3 Sample

4.3.1 Final Attributes

Once the data has been cleaned, is it necessary to assign names to the different variables selected, which are going to be the attributes of the model. As we have mentioned before, the data obtained is from the last and last but one accounts registered. Although some variables do not need a variable for each year, there are some that yes, as we can see in Table 4.1. The final attributes of the model we are going to validate are the following:

	Last year	Last year but one
365/Stock Rotation	365/R	365/R1
Collection Period	TCLIENTS	TCLIENTS1
Suppliers Payment Period	TPROV	TPROV1
Share Capital	CSOC	CSOC
Legal Form	FJUR	FJUR
Number of Companies in Corporate Group	NCOMP	NCOMP
Audit	AUDIT	AUDIT1
CNAE	CNAE	CNAE
Op.Result/Op.Income	REXP/INGEXP	REXP/INGEXP1
Sales/Total Assets	V/AT	V/AT1
Current Assets/Current Liabilities.	AC/PC	AC/PC1
Current Assets/Total Assets	AC/AT	AC/AT1
CurRent Liabilities/Total Liabilities	PC/PT X	PC/PT1
Financial Debt/Cash Flow	DF/CF	DF/CF1
Total Assets/Total Debt	AT/DT	AT/DT1
Sales evolution	EVOV	EVOV1
Employees evolution	EVOE	EVOE1
Size	TAMANY	TAMANY1
Age	EDAT	EDAT

Table 4.1: Final variables of the model
Source: Own elaboration

The target field of the dataset will be a variable named TARGET, which will be binary with "YES" or "NO" values. The companies that have failed will receive the value "YES" and the other companies the value "NO".

4.3.2 Final Sample

Finally, we obtained information about 1208 failed companies and 4410 from non-failed companies. In Table 4.2, we can see a summary of the results obtained from the final data. The different values are the mean calculated from all the entries.

As we can see in Table 4.2, the structure of the companies is similar, according to the way their assets and liabilities are divided. The companies of the construction sector tend to have more current assets than long-term, because it is something usual to rent the heavy machinery, avoiding oversize the long-term assets of the company. According to the summary, the ratio AC/AT oscillates between 0,64 and 0,71. The ratio PC/PT also shows the similarities of the liabilities structure, where it oscillates between 0,62-0,71. This means that never mind the company has failed or not, the assets and liabilities structure is similar.

Where we first start appreciating differences between failed and non-failed companies are in TProv, that is the time that passes between we buy a product and we pay it. In the companies that fail, the value increments significantly between the year before and the last year,which clearly shows us that the company is struggling to pay their suppliers. In the healthy companies, the mean stays the same.

Another value that differs according to the type of company is CSOC. Active companies have

higher share capital than the ones that have failed.

The efficiency also varies as we can see in the results of the ratio RENDEXP/INGEXP. The failed companies have negative efficiency, and it also decreases significantly the last year. However, healthy companies have a positive efficiency.

Talking about debt and the ratio AT/DT, we can see that active companies have more resources to face the debt they have. We can also see a big difference in the ratio DF/CF, where the failed companies have more problems paying their financial debts with their cash flow than the active ones. It also increases a lot during the last year before the bankruptcy.

Finally, the evolution of the number of employees and sales distinguishes clearly the two type of companies. While in the active companies the mean of the evolution is positive, in the failed companies is negative.

	CONCURSADES		**NO CONCURSADES**	
	ANY	**ANY -1**	**ANY**	**ANY -1**
365/ROTACIÓ	439,58	278,39	690,67	702,37
T.CLIENTS	134,97	123,58	117,58	124,95
T.PROVEIDORS	108,08	80,79	66,09	70,00
CSOC	516,28		2962,88	
NCOMP	0,81		0,84	
RENDEXP/INGEXP	-0,29	-0,08	0,05	0,05
V/AT	1,15	1,23	0,77	0,76
AC/PC	1,23	1,47	3,26	3,28
AC/AT	0,67	0,71	0,64	0,65
PC/PT	0,71	0,71	0,62	0,62
AT/DT	1,09	1,26	1,85	1,85
DF/CF	670,14	406,21	152,85	151,82
EVOV	-13%	-11%	75%	21%
EVOE	-6%	-8%	10%	12%
TAMANY	3,33	3,84	3,5	3,65
EDAT	23,95		26,8	

Table 4.2: Sample summary
Source: Own elaboration

For the non-categorical variables, the results obtained are the following:

- **Legal form:** In the group of failed companies, 80% is represented by private limited companies , 18,7% by public limited companies, and 1,3% by cooperatives. The same results were obtained from the non-failed companies.

- **Audit:**: In this field, not all companies have been audited, so the results are based on the ones that have been audited. From the failed companies, 65% have negative comments, 12,5% have favorable with uncertainties comments, and 22,5% have positive comments on the last year. 2 years before, the results were 53,8% negatives, 6,7% favorable with uncertainties, and 39,5% positives. In this variable, we can see an aggravation of the negative results from 2 years before to the last year, and a reduction of the positive ones.

On the other hand, the non-failed companies' results the year before were 49,25% positive, 23,6% favorable with uncertainties, and 27,15% negatives. With the 2 years before data the results were 65% positives, 7% favorable with uncertainties, and 28% negatives. With these results, we can see a significant difference between the two samples.

- **CNAE:** The main specific type of companies in the samples are residential buildings construction, electrical constructions, and building promotion, having the 1st one weight of over 30%. If we compare it with the construction market, we can see that residential building construction is the one that has more weight, too.

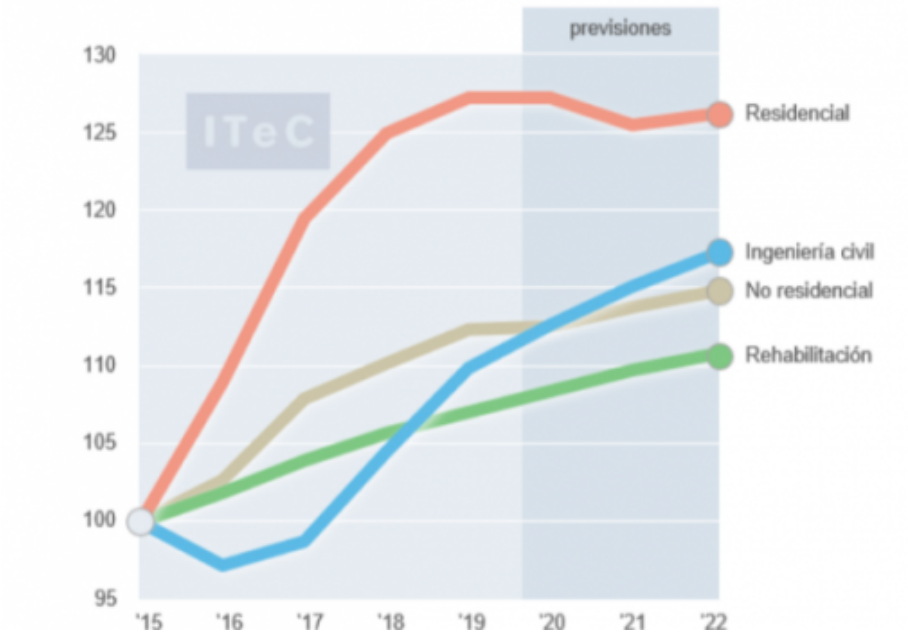

Figure 4.1: Evolution of construction subsectors

Source: Euroconstruct, I. (2019). Estudio de mercado de la construcción euroconstruct – diciembre 2019.

Chapter 5

Experimental Evaluation

This chapter aims to explain how we have applied Machine Learning in this project and to obtain the final results of the hypothesis explained in Chapter 1.

5.1 Software used

5.1.1 Pandas

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

5.1.2 Numpy

NumPy is a Python library used for working with arrays. It also has functions for working in the domain of linear algebra, Fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open-source project and you can use it freely. NumPy stands for Numerical Python. NumPy contains a multi-dimensional array and matrix data structures. It can be utilised to perform several mathematical operations on arrays such as trigonometric, statistical, and algebraic routines. Pandas objects rely heavily on NumPy objects.

5.1.3 Scikit-learn

Scikit-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbours, and it also supports Python numerical and scientific libraries like NumPy and SciPy. Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering, and dimensional reduction via a consistent interface in Python.

We have used this library to process the data, apply different algorithms to it and validate them.

5.2 Preprocessing

Once we have the two data sets in .csv format, we need to process this data in order to apply the algorithms. To accomplish it, we have used the programming language Python and its libraries

Pandas, Numpy and Sklearn.

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its wide range of libraries and packages allows working efficiently. The application used for the programming code has been Visual Studio Code.

Firstly, we have applied one-hot-encoding for categorical features, which encoded categorical features as a one-hot numeric array. This encoding is needed for feeding categorical data to many scikit-learn estimators, notably linear models and SVMs with the standard kernels.

Secondly, to balance the classes and having the same number of companies in both samples, we reduced the healthy companies sample to 1208.

Learning the parameters of a prediction function and testing it on the same data is a methodological mistake: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data. This situation is called overfitting. A solution to this problem is a procedure called cross-validation. A test set should still be held out for final evaluation, but the validation set is no longer needed when applying cross-validation. In the basic approach, called k-fold cross-validation, the training set is split into k smaller sets. The following procedure is followed for each of the k "folds":

- A model is trained using the folds as training data;
- the resulting model is validated on the remaining part of the data (it is used as a test set to compute a performance measure such as accuracy).

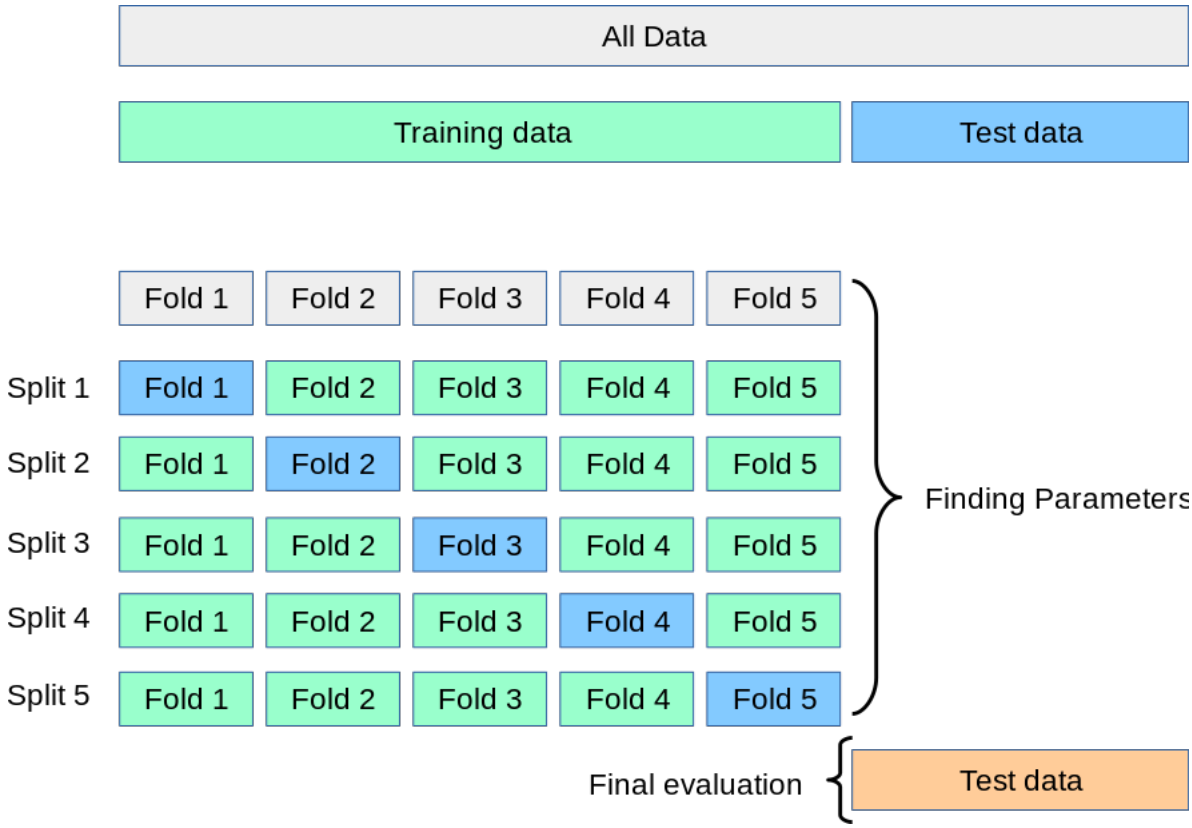

Figure 5.1: Cross-validation process
Source: Scikit-learn User Guide

The performance measure reported by k-fold cross-validation is then the average of the values

computed in the loop. This approach can be computationally expensive but does not waste too much data (as is the case when fixing an arbitrary validation set), which is a major advantage in problems such as inverse inference where the number of samples is very small.

5.3 First approach on validating the model

5.3.1 Algorithms used

The estimators used in Scikit-learn library are all supervised learning algorithms. Supervised learning, also known as supervised machine learning, is a subcategory of machine learning and artificial intelligence. It is defined by its use of labeled datasets to train algorithms that classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately, which occurs as part of the cross-validation process.

- **Decision Tree:** Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation. DecisionTreeClassifier is a class capable of performing multi-class classification on a dataset.

  - **DecisionTreeClassifier:** As with other classifiers, DecisionTreeClassifier takes as input two arrays: an array X, sparse or dense, of shape (n-samples, n-features) holding the training samples, and an array Y of integer values, shape (n-samples,), holding the class labels for the training samples. In case that there are multiple classes with the same and highest probability, the classifier will predict the class with the lowest index amongst those classes.

- **Ensemble methods:** The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm to improve generalizability / robustness over a single estimator. Two families of ensemble methods are usually distinguished:

  In averaging methods, the driving principle is to build several estimators independently and then to average their predictions. On average, the combined estimator is usually better than any of the single base estimators because its variance is reduced.

  By contrast, in boosting methods, base estimators are built sequentially and one tries to reduce the bias of the combined estimator. The motivation is to combine several weak models to produce a powerful ensemble.

  - **AdaBoostClassifier:** Situated in the boosting ensemble methods. The core principle of AdaBoost is to fit a sequence of weak learners (models that are only slightly better than random guessings, such as small decision trees) on repeatedly modified versions of the data. The predictions from all of them are then combined through a weighted majority vote (or sum) to produce the final prediction. The data modifications at each so-called boosting iteration consist of applying weights to each of the training samples. Initially, those weights are all set to 1/N, so that the first step simply trains a weak learner on the original data. For each successive iteration, the sample weights are individually modified and the learning algorithm is reapplied to the reweighted

data. At a given step, those training examples that were incorrectly predicted by the boosted model induced at the previous step have their weights increased, whereas the weights are decreased for those that were predicted correctly. As iterations proceed, examples that are difficult to predict receive ever-increasing influence. Each subsequent weak learner is thereby forced to concentrate on the examples that are missed by the previous ones in the sequence.

  - **GradientBoostingClassifier:** Situated in the boosting ensemble methods. Gradient Boosting builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage, n-classes regression trees are fit on the negative gradient of the binomial or multinomial deviance loss function. Binary classification is a special case where only a single regression tree is induced.
  - **RandomForestClassifier:** Situated in the average ensemble methods. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max-samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

- **Support Vector Machine:** Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression, and outliers detection. The advantages of support vector machines are:

  - Effective in high dimensional spaces.
  - Still effective in cases where a number of dimensions are greater than the number of samples.
  - Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
  - Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

  - **SVC:** Support Vector Classification is a class capable of performing binary and multiclass classification on a dataset. As other classifiers, SVC takes as input two arrays: an array X of shape (n-samples, n-features) holding the training samples, and an array y of class labels (strings or integers), of shape (n-samples):

Once the data has been processed, it is time to prove which algorithm offers better results.

Listing 5.1 shows the code used in this project to estimate the model. As we can see in lines 37 to 41, we have applied the one-hot-encoding process mentioned in Section 5.2. From line 47 to the end, we have applied cross-validation mentioned also in Section 5.2, selecting the different estimators.

```
1 import pandas as pd
2 import numpy as np
3 from sklearn import tree
4 from sklearn.tree import DecisionTreeClassifier
```

```
5  from sklearn.ensemble import AdaBoostClassifier,
       GradientBoostingClassifier, RandomForestClassifier
6  from sklearn.svm import SVC
7
8  from sklearn.preprocessing import OneHotEncoder
9  from sklearn.model_selection import cross_val_predict
10 from sklearn.metrics import accuracy_score, precision_score,
       recall_score, f1_score, confusion_matrix
11
12 def load_data(path):
13     df = pd.read_csv(
14         path,
15         sep=";",
16         thousands=".",
17         decimal=",",
18         encoding="ISO-8859-1",
19     )
20     df = df.dropna(axis=0, how="all")
21     df["EVOV"] = df["EVOV"].str.replace("%", "").astype(float)
22     df["EVOE"] = df["EVOE"].str.replace("%", "").astype(float)
23     return df
24
25
26 # Load & concat data
27 df_y = load_data("dataset-concursades.csv")
28 df_n = load_data("dataset-no-concursades.csv")
29 df_n = df_n.sample(1208)
30 df = pd.concat([df_y, df_n], axis=0)
31
32 # Split X and y
33 X = df.drop("TARGET", axis=1)
34 y = df["TARGET"]
35
36 # One-hot-encoding of categorical features
37 ohe = OneHotEncoder(sparse=False)
38 X_categorical = X.loc[:, X.columns[X.dtypes == "object"]]
39 X_numerical = X.loc[:, X.columns[X.dtypes != "object"]]
40 X_ohe = ohe.fit_transform(X_categorical)
41 X_preproc = np.concatenate([X_numerical.values, X_ohe], axis=1)
42
43 # Label encoding of y (YES=1, NO=0)
44 y_preproc = (y == "YES").astype(int)
45
46 # Cross-validate estimators
47 seed = 1234
48 estimators = [
```

```
    DecisionTreeClassifier(random_state=seed),
    AdaBoostClassifier(random_state=seed),
    GradientBoostingClassifier(random_state=seed),
    RandomForestClassifier(random_state=seed),
    SVC(random_state=seed),
]
for estim in estimators:
    y_pred = cross_val_predict(estim, X_preproc, y_preproc, cv=10)
    print(f"***** {estim.__class__.__name__} *****")
    print(f"Accuracy: {accuracy_score(y_preproc, y_pred)}")
    print(f"Precision (positive class 'YES'): {precision_score(
        y_preproc, y_pred)}")
    print(f"Recall (positive class 'YES'): {recall_score(y_preproc,
         y_pred)}")
    print(f"f1-score (positive class 'YES'): {f1_score(y_preproc,
        y_pred)}")

    print("Confusion matrix:")
    print(confusion_matrix(y_preproc, y_pred), end="\n\n")
```

Listing 5.1: Implementation of evaluation process

5.4 Final Results

The first step when all the data has been treated and ready to be executed was to see the general estimates of the model with all the data gathered from the last two years. The estimators used have been:

- **Accuracy:**In multilabel classification, this function computes subset accuracy: the set of labels predicted for a sample must exactly match the corresponding set of labels in y-true.
- **Precision:**The precision is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative. The best value is 1 and the worst value is 0.
- **Recall:**The recall is the ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples. The best value is 1 and the worst value is 0.
- **F1-score:**The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is:

  F1 = 2 * (precision * recall) / (precision + recall)

Once we have obtained the general results of the study, it is important to analyse them in a more precise way. We estimated the model with the data from one year before and two years

before separately, to see if the results plunged during this period, something usual in previous studies.

As we can see in the following listing, the results show high percentage estimations, obtaining from AdaBoostClassifier and GradientBoostingClassifier algorithms over 98% rate. In this case, Support Vector Classification (SVM) has a more reduced accuracy in comparison with the other algorithms used. We also can observe that Precision gives slightly better results than Accuracy and the other ways of evaluating the model.

```
1
2   ***** DecisionTreeClassifier *****
3  Accuracy: 0.9838509316770186
4  Precision (positive class 'YES'): 0.9858569051580699
5  Recall (positive class 'YES'): 0.9817729908864954
6  f1-score (positive class 'YES'): 0.983810709838107
7  Confusion matrix:
8  [[1191   17]
9   [  22 1185]]
10
11 ***** AdaBoostClassifier *****
12 Accuracy: 0.9813664596273292
13 Precision (positive class 'YES'): 0.9849749582637729
14 Recall (positive class 'YES'): 0.9776304888152444
15 f1-score (positive class 'YES'): 0.9812889812889812
16 Confusion matrix:
17 [[1190   18]
18  [  27 1180]]
19
20 ***** GradientBoostingClassifier *****
21 Accuracy: 0.9809523809523809
22 Precision (positive class 'YES'): 0.9874055415617129
23 Recall (positive class 'YES'): 0.9743164871582436
24 f1-score (positive class 'YES'): 0.9808173477898249
25 Confusion matrix:
26 [[1193   15]
27  [  31 1176]]
28
29 ***** RandomForestClassifier *****
30 Accuracy: 0.9569358178053831
31 Precision (positive class 'YES'): 0.9669771380186283
32 Recall (positive class 'YES'): 0.9461474730737366
33 f1-score (positive class 'YES'): 0.9564489112227805
34 Confusion matrix:
35 [[1169   39]
36  [  65 1142]]
37
```

```
***** SVC *****
Accuracy: 0.7523809523809524
Precision (positive class 'YES'): 0.8327868852459016
Recall (positive class 'YES'): 0.6313173156586578
f1-score (positive class 'YES'): 0.7181903864278981
Confusion matrix:
[[1055  153]
 [ 445  762]]
```

Listing 5.2: Global Results

Continuing with the process of estimating the capacity of the model to predict insolvency, we have analysed the last years' data, obtaining the following results:

```

***** DecisionTreeClassifier *****
Accuracy: 0.9780538302277433
Precision (positive class 'YES'): 0.9816360601001669
Recall (positive class 'YES'): 0.9743164871582436
f1-score (positive class 'YES'): 0.977962577962578
Confusion matrix:
[[1186   22]
 [  31 1176]]

***** AdaBoostClassifier *****
Accuracy: 0.9846790890269151
Precision (positive class 'YES'): 0.9866888519134775
Recall (positive class 'YES'): 0.9826014913007457
f1-score (positive class 'YES'): 0.9846409298464094
Confusion matrix:
[[1192   16]
 [  21 1186]]

***** GradientBoostingClassifier *****
Accuracy: 0.9830227743271222
Precision (positive class 'YES'): 0.9915682967959528
Recall (positive class 'YES'): 0.9743164871582436
f1-score (positive class 'YES'): 0.9828666945257
Confusion matrix:
[[1198   10]
 [  31 1176]]

***** RandomForestClassifier *****
Accuracy: 0.9627329192546584
Precision (positive class 'YES'): 0.970513900589722
Recall (positive class 'YES'): 0.9544324772162386
f1-score (positive class 'YES'): 0.962406015037594
```

```
Confusion matrix:
[[1173   35]
 [  55 1152]]

***** SVC *****
Accuracy: 0.7677018633540372
Precision (positive class 'YES'): 0.8533916849015317
Recall (positive class 'YES'): 0.6462303231151616
f1-score (positive class 'YES'): 0.7355021216407356
Confusion matrix:
[[1074  134]
 [ 427  780]]
```

Listing 5.3: Results 1 year before

The results offer a capacity of prediction between 95% and 99% if we exclude SVC.

As the results show a very high capacity of prediction, we have repeated the execution 1000 times to prove that the they continue in the same line, and we have obtained the following:

```

***** DecisionTreeClassifier *****
Accuracy: 0.9770567287784694
Precision (positive class 'YES'): 0.9789112830402317
Recall (positive class 'YES'): 0.9751085335542673
f1-score (positive class 'YES'): 0.9769880495208504


***** AdaBoostClassifier *****
Accuracy: 0.9829780538302282
Precision (positive class 'YES'):  0.98453156475216
Recall (positive class 'YES'): 0.981367025683513
f1-score (positive class 'YES'): 0.9829430348584495


***** GradientBoostingClassifier *****
Accuracy: 0.9802939958592134
Precision (positive class 'YES'): 0.9855903340953686
Recall (positive class 'YES'):  0.974830157415079
f1-score (positive class 'YES'): 0.9801781676012492

***** RandomTreeClassifier *****
Accuracy: 0.964392958562564
Precision (positive class 'YES'): 0.970724915743686
Recall (positive class 'YES'):  0.9576 30515410759
f1-score (positive class 'YES'): 0.9641554780152234
```

Listing 5.4: Results 1 year before analysed

As we can see, after repeating the operation the results fall slightly, but it continues with a capacity of prediction over 98%.

Finally, if we analyse the data from 2 years before, we can see that the results are about 97% rate, which means that the model is strong and does not plunge its capacity of prediction 2 years before the insolvency.

```
1
2  ***** DecisionTreeClassifier *****
3  Accuracy: 0.9755693581780538
4  Precision (positive class 'YES'): 0.9743801652892562
5  Recall (positive class 'YES'): 0.9768019884009942
6  f1-score (positive class 'YES'): 0.9755895738518825
7  Confusion matrix:
8  [[1177   31]
9   [  28 1179]]
10
11 ***** AdaBoostClassifier *****
12 Accuracy: 0.9768115942028985
13 Precision (positive class 'YES'): 0.9775933609958506
14 Recall (positive class 'YES'): 0.975973487986744
15 f1-score (positive class 'YES'): 0.9767827529021559
16 Confusion matrix:
17 [[1181   27]
18  [  29 1178]]
19
20 ***** GradientBoostingClassifier *****
21 Accuracy: 0.979296066252588
22 Precision (positive class 'YES'): 0.9777043765483072
23 Recall (positive class 'YES'): 0.9809444904722452
24 f1-score (positive class 'YES'): 0.9793217535153019
25 Confusion matrix:
26 [[1181   27]
27  [  23 1184]]
28
29 ***** RandomForestClassifier *****
30 Accuracy: 0.9685300207039338
31 Precision (positive class 'YES'): 0.975609756097561
32 Recall (positive class 'YES'): 0.9610604805302403
33 f1-score (positive class 'YES'): 0.9682804674457429
34 Confusion matrix:
35 [[1179   29]
36  [  47 1160]]
37
38 ***** SVC *****
39 Accuracy: 0.5412008281573499
```

```
Precision (positive class 'YES'): 0.5242528172464478
Recall (positive class 'YES'): 0.8864954432477217
f1-score (positive class 'YES'): 0.6588669950738916
Confusion matrix:
[[ 237  971]
 [ 137 1070]]
```

Listing 5.5: Results 2 years before

If we repeat the execution 1000 times, the algorithms that show better results are the following:

```

[style=PyScript,frame=single,caption={Implementation of evaluation
   process}, captionpos=b, label={lst:evaluation-code}]

***** DecisionTreeClassifier *****
Accuracy: 0.9743378881987587
Precision (positive class 'YES'): 0.9747528693579689
Recall (positive class 'YES'): 0.9738889809444903
f1-score (positive class 'YES'): 0.9743146981750355

***** AdaBoostClassifier *****
Accuracy: 0.9782368530020719
Precision (positive class 'YES'): 0.9782050429984809
Recall (positive class 'YES'): 0.9782584921292461
f1-score (positive class 'YES'): 0.978226939609969

***** GradientBoostingXlassifier *****
Accuracy: 0.978578053830229
Precision (positive class 'YES'): 0.9768246716422597
Recall (positive class 'YES'): 0.9804059652029817
f1-score (positive class 'YES'): 0.9786084062342008

***** RandomTreeClassifier *****
Accuracy: 0.965592958560221
Precision (positive class 'YES'): 0.968143765743686
Recall (positive class 'YES'):  0.9627 30414560759
f1-score (positive class 'YES'): 0.9654 554780152234
```

Listing 5.6: Results 2 years before analysed

We can confirm that the results after executing repeated times the script, continue with an over 97% rate.

To end the experimental evaluation chapter, we can conclude that the model we have created works with data from the year before and from the year before but one, obtaining results between 95% and 98% capacity of prediction.

In Tables 5.1 and 5.2, we can see the results with the data of the last year and last year but one resumed, highlighting the best results. As we can see, the algorithms that have offered better results are AdaBoostClassifier and GradientBoostingClassifier.

	Accuracy	Precision	F-Score	Recall
DecisionTreeClassifier	0.9770	0.9789	0.9751	0.9769
AdaBoostClassifier	0.9829	0.9845	0.9813	0.9829
RandomTreeClassifier	0.9643	0.9707	0.9576	0.9641
GradientBoostingClassifier	0.9802	0.9855	0.9748	0.9801

Table 5.1: Final results resume year before
Source: Own elaboration

	Accuracy	Precision	F-Score	Recall
DecisionTreeClassifier	0.9743	0.9747	0.9738	0.9743
AdaBoostClassifier	0.9780	0.9783	0.9783	0.9781
RandomTreeClassifier	0.9655	0.9681	0.9627	0.9654
GradientBoostingClassifier	0.9785	0.9768	0.9804	0.9786

Table 5.2: Final results resume year before
Source: Own elaboration

Chapter 6

Conclusions and future work

In this project, we have confirmed our hypothesis that is possible to predict the failure of a company having its last two years' annual accounts with the use of Machine Learning. In particular, it is possible to predict insolvency with the data we have obtained of 1 year before with a rate of over 98%, and a rate of over 97% with the data of 2 years before. The fact that the results do not plunge determines the consistency of the model, as many previous studies failed when trying the model extending it to more than one year. Compared with other failure prediction studies results, this one has obtained high levels of rating.

As future work, and in relation with the situation that many companies are struggling and will continue during the next years, a list of companies that need to change if they do not want to fail could be created. Moreover, the contribution of each ratio to offer good results to the model could be analysed.

Concerning my experience, during this 8-month period, I have learned a lot about a problem that many companies are suffering, which is bankruptcy. My motivation on this topic has been increasing during these months, as the goal of the project was to offer a solution to a real problem. Moreover, I have experienced how it is to work with the Python library Scikit-learn and the supervised learning algorithms it offers.

Bibliography

[Abid and Zouari, 2000] Abid, F. and Zouari, A. (2000). Financial distress prediction using neural networks. *SSRN Electronic Journal*.

[Abidali and F.Harris, 1995] Abidali, A. and F.Harris (1995). A methodology for predicting company failure in the construction industry. *Research Gate*.

[Altman, 1968] Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*.

[Altman and Sabato, 2007] Altman, E. and Sabato, G. (2007). Modelling credit risk for smes: Evidence from the u.s. market. *Abacus*, 43:332–357.

[Altman, 2015] Altman, E. I. (2015). Financial and non-financial variables as long-horizon predictors of bankruptcy. *Social Science Research Network*.

[Balcaen and Ooghe, 2006] Balcaen, S. and Ooghe, H. (2006). 35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems. *The British Accounting Review*, 38:63–93.

[Beaver, 1966] Beaver, W. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*.

[Chan et al., 2005] Chan, J., Tam, C., and Cheung, R. (2005). Construction firms at the crossroads in hong kong: Going insolvency or seeking opportunity. *Engineering, Construction and Architectural Management*, 12:111–124.

[Davydenko and Franks, 2008] Davydenko, S. A. and Franks, J. R. (2008). Do bankruptcy codes matter? a study of defaults in france, germany, and the u.k. *The Journal of Finance*, 63(2):565–608.

[du Jardin, 2009] du Jardin, P. (2009). Bankruptcy prediction models: How to choose the most relevant variables. *Edhec Business School*.

[Fernández-Gámez et al., 2016] Fernández-Gámez, M., Cisneros-Ruiz, A., and Callejon, A. (2016). Applying a probabilistic neural network to hotel bankruptcy prediction. *Tourism and Management Studies*, 12:40–52.

[Fletcher and Goss, 1993] Fletcher, D. and Goss, E. (1993). Forecasting with neural networks: An application using bankruptcy data. *Information and Management*, 24(3):159–167.

[Gil de Albornoz and B.Giner, 2013] Gil de Albornoz, B. and B.Giner (2013). Predicción del fracaso empresarial en los sectores construcción e inmobiliario: modelos generales versus específicos. *Universia Business Review*.

[González and y A. Y. Sánchez, 2003] González, J. P. and y A. Y. Sánchez, J. C. (2003). El reglamento sobre procedimientos de insolvencia de la unión europea. *Partida Doble,141*.

[Grunert et al., 2005] Grunert, J., Norden, L., and Weber, M. (2005). The role of non-financial factors in internal credit ratings. *Journal of Banking and Finance*, 29(2):509–531.

[INE, 2019] INE (2019). Estructura y dinamismo del tejido empresarial en españa directorio central de empresas (dirce) a 1 de enero de 2019. *Instituto Nacional de Estadística*.

[Jodi L. Bellovary and Akers, 2007] Jodi L. Bellovary, D. E. G. and Akers, M. D. (2007). A review of bankruptcy prediction studies: 1930-present. *Journal of Financial Education*.

[Karels and Prakash, 2006] Karels, G. and Prakash, A. (2006). Multivariate normality and forecasting of business bankruptcy. *Journal of Business Finance and Accounting*, 14:573 – 593.

[Keasey and Watson, 2006] Keasey, K. and Watson, R. (2006). Non-financial symptoms and the prediction of small company failure: A test of argenti's hypotheses. *Journal of Business Finance and Accounting*, 14:335 – 354.

[Laguillo et al., 2017] Laguillo, G., del Castillo, A., and Manuel Ángel Fernández, R. B. (2017). Modelos centrados vs descentrados para la predicción de quiebra: evidencia empírica para españa. *DialNet*.

[Laitinen and Kankaanpaa, 1999] Laitinen, T. and Kankaanpaa, M. (1999). Comparative analysis of failure prediction methods: the Finnish case. *European Accounting Review*, 8(1):67–92.

[Manuel Rodríguez López and de Llano Monelos, 2006] Manuel Rodríguez López, C. P. S. and de Llano Monelos, P. (2006). Predicción de insolvencia y fracaso financiero: medio siglo después de beaver(1966). avances y nuevos resultados. *Finance and Management Information Systems Research Group (FYSIG)*.

[Mateos-Ronco and Mas, 2011] Mateos-Ronco and Mas, A. (2011). Developing a business failure prediction model for cooperatives: Results of an empirical study in spain. *African journal of business management*, 5:10565–10576.

[Min and Lee, 2005] Min, J. and Lee, Y.-C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications*, 28:603–614.

[Mínguez-Conde, 2006] Mínguez-Conde (2006). La información contable en la empresa constructora: factores identificativos del fracaso empresarial. *Universidad de Valladolid*.

[Pang, 2013] Pang (2013). Retail bankruptcy prediction. *American Journal of Economics and Business Administration*, 5:29–46.

[Park and Hancer, 2012] Park, S.-S. and Hancer, M. (2012). A comparative study of logit and artificial neural networks in predicting bankruptcy in the hospitality industry. *Tourism Economics*, 18:311–338.

[Peres and Antão, 2017] Peres, C. and Antão, M. (2017). The use of multivariate discriminant analysis to predict corporate bankruptcy: A review. *AESTIMATIO, THE IEB INTERNATIONAL JOURNAL OF FINANCE*, page 108.

[Santomero and Vinso, 1977] Santomero, A. M. and Vinso, J. D. (1977). Estimating the probability of failure for commercial banks and the banking system. *Journal of Banking and Finance*, 1(2):185–205.

[Serrano-Cinca and Martín-del Brío, 1993] Serrano-Cinca, C. and Martín-del Brío, B. (1993). Predicción de la quiebra bancaria mediante el empleo de redes neuronales artificiales. *Revista española de financiación y contabilidad, ISSN 0210-2412, Nº 74, 1993, pags. 153-176*, 22.

[Shi and Li, 2019] Shi, Y. and Li, X. (2019). An overview of bankruptcy prediction models for corporate firms: A systematic literature review. *Intangible Capital*, 15:114.

[Shin et al., 2005] Shin, K.-s., Lee, T., and Kim, H.-j. (2005). Shin, k.s.: An application of support vector machines in bankruptcy prediction model. expert systems and applications 28, 127-135. *Expert Systems with Applications*, 28:127–135.

[Stroe and Bărbuţă-Mişu, 2010] Stroe, R. and Bărbuţă-Mişu, N. (2010). Predicting the financial performance of the building sector enterprises – case study of galati county (romania). *The Review of Finance and Banking*, 02:029–039.

[Tam and Kiang, 1992] Tam, K. Y. and Kiang, M. Y. (1992). Managerial applications of neural networks: The case of bank failure predictions. *Management Science*, 38(7):926–947.

[Wang et al., 2004] Wang, Y., Lo, H., Chi, R., and Yang, Y. (2004). An integrated framework for customer value and customer-relationship-management performance: A customer-based perspective from china. *Managing Service Quality*, 14:169–182.

[Wilson, 2013] Wilson, L. (2013). Family business survival and the role of boards. *Entrepreneurship Theory and Practice*, 37.

[Wilson and Sharda, 1994] Wilson, R. L. and Sharda, R. (1994). Bankruptcy prediction using neural networks. *Decision Support Systems*, 11(5):545–557.

[Zavgren, 2006] Zavgren, C. (2006). Assessing the vulnerability to failure of american industrial firms: A logistic analysis. *Journal of Business Finance and Accounting*, 12:19 – 45.

[Zavgren and Friedman, 1987] Zavgren, C. V. and Friedman, G. E. (1987). Are bankruptcy prediction models worthwhile? an application in securities analysis. *Management International Review*, 28(1):34–44.